新媒体数据分析与应用浅析

王楷鑫[1] 郗 睿[1] 王立国[2] 王双立[2*]
（1. 山西传媒学院，山西 晋中 030619；2. 北华大学，吉林 吉林 132021）

摘 要：以数据分析需求的视角描述了新媒体数据采集目标的设定及采集方式；在介绍以促进质量与效率为目的的新媒体数据预处理之后，以实例化方式着力阐述了矩阵分析法、相关性分析法及回归分析法为代表的统计分析和 BP 神经网络为代表的算法模型等经典分析方法的概念、原理及实现方法；讲述针对海量数据的大数据分析功能及基于 Hadoop 的 KNN 分类算法的应用设计，为新媒体运营企业从容应对新媒体数据分析提供了理论依据和技术支撑。

关键词：新媒体数据分析；BP 神经网络；大数据分析；KNN 分类算法　**中图分类号：**TN948.1　**文献标志码：**A
文章编号：1671-0134（2021）10-145-04　DOI：10.19483/j.cnki.11-4653/n.2021.10.044
本文著录格式：王楷鑫，郗睿，王立国，王双立 . 新媒体数据分析与应用浅析 [J]. 中国传媒科技，2021（10）：145-148.

导语

新媒体时代背景下，外部信息除传统意义上需要便捷快速获取外还被附加了高效处理与分析、精准投放、服务运营等要求，这对国家、企业乃至个人至关重要。新媒体与云计算、大数据、人工智能等现代技术的深度融合为社会提供了更加优质的数据应用及智能服务，为新媒体运营企业找准方向、降低成本、规划议案提供了可靠依据。

1. 新媒体数据采集

1.1 设定采集目标

在当前新媒体数据呈现海量的情况下，采集前必须依据数据分析的需求来界定采集目标、设置采集范围、排除冗余数据以增强数据的代表性与可信度。从现实问题中找出解决问题的关键节点，提取相关事务的特征属性，依据特征属性规划数据分析方向、提炼采集目标。

1.2 数据来源及采集

新媒体数据一般是在社会生产、管理、运营过程中产生的，因而主要来源于网络数据库、社交媒体、网络舆情及系统运行日志等方面。新媒体数据采集本质上是依据新媒体数据来源运用多种方式收集，一般使用运营方（或管理方）的数据库及第三方平台数据两种方式。两者常见于从运营系统的服务器直接获取和网络中使用智能爬虫技术进行云端采集，达到数据实时汇集的目标。此外手工问卷调查作为前两者的补充有利于调查者与受访者现场沟通、精准把握受访者的心理特征，从而明确受访者的需求。

2. 新媒体数据预处理

数据预处理是指在数据主要处理与分析前进行的加工整理，达到清理异常、纠正错误、统一格式等目标，使用数据清洗、数据集成、数据变换等方法提高了数据分析的质量与效率。其中的数据清洗现已成为大数据预处理的常用方法，主要实现删除重复信息、纠正存在的错误，提供数据一致性等目标。

3. 新媒体数据分析

新媒体数据分析是指用适当的分析方法对大量新媒体数据进行剖析与加工，使其易于理解并反映数据信息所代表的现实事物本质特征及内在规律，以最大限度地发挥数据的作用。常用经典的分析方法有操作简洁的统计分析、高度复杂的算法模型及针对海量数据的大数据分析。

3.1 统计分析

3.1.1 矩阵分析法

矩阵分析法是以待分析数据的两个重要指标作为横、纵坐标轴构成四个象限来分析问题，提出解决问题的合理方法并汲取数据分析结论。以 Kano 模型为代表的矩阵分析法如图 1 所示。

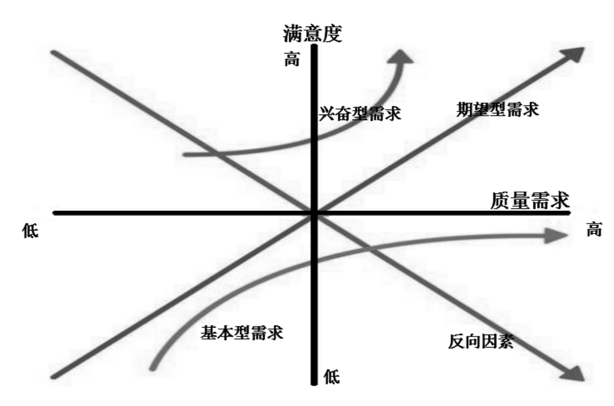


图 1　矩阵分析法 Kano 模型

基金项目：山西省教育科学“十三五”规划项目，项目编号：GH-19125。吉林省吉林市社科联项目，项目编号：19101。

* 为通讯作者

模型中的兴奋型需求是用户完全意想不到的或处于潜意识状态中的、需要挖掘与洞察的需求，当此需求被提供后，用户会产生意外惊喜进而表现出非常满意，若不能提供则满意度会下降。期望型需求作为一维因素与满意度成正比，它是成长期的需求，客户、竞争对手和运营企业自身都需关注的需求，它体现了竞争能力，运营企业应注重提高此类需求的服务质量。基本型需求是用户对产品的必备需求，要求服务产品必须具有相关功能；当不断强化产品功能后，用户满意度不会显著提升，若消除此产品功能，用户满意度将明显下降。作为用户完全不需要的反向因素与用户满意度成反比。所以在设计产品期间，尽量避免反向因素的出现，做好基本型需求、不断完善期望型需求的质量，突出兴奋型需求。

Kano 模型调研的每个功能需求都有正向和负向两种评价，依据每种功能的需求可按喜欢、理应如此、勉强接受、我不喜欢四个值进行评价形成二维表，折算 Better-Worse 系数，Better 系数表示满意系数，常为正值，Worse 系数表示为不满意系数，常为负值。图 2 是 Better-Worse 系数对应的需求分析，可见第一象限内产品功能 1 是期望型需求最优的，可以优先做。

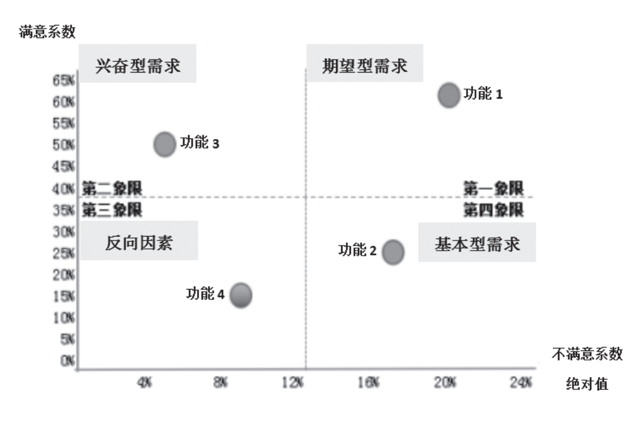


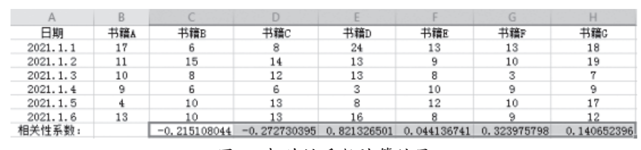
图 2 Better-Worse 系数对应的需求分析

Kano 模型需要结合业务的本身特征来构建，否则得出的分析结论可能与实际情况存在较大偏差，这就需要问卷调查设计的问题能够精准反映产品的特性及寻找合适的问卷服务对象。另外需要注意的是，某种类型的需求会随着时间的推移而演变成另一种类型的需求。因此，需要持续调研需求，并依据数据分析结果更新产品特性。

3.1.2 相关性分析法

相关性分析是用来衡量两个或多个变量因素间相关密切程度的，相关性不等价于因果关系。在新媒体营销中，可以通过比较两个商品的相关关系强弱来选择是否进行组合销售。[1]

图 3 是运用 Excel 中的 CORREL 函数对在 2021 年 1 月 1 日至 6 日期间书籍 A 销售数量与书籍 B 至书籍 G 销售数量相关性系数计算结果。相关性系数取值范围为 [−1, 1]，其绝对值越大，相关程度越大，由计算结果可知，书籍 A 与书籍 D 的相关性系数最高，为 0.821326501，因此两者搭售可激发用户更多的购买行为。



A	B	C	D	E	F	G	H
日期	书籍A	书籍B	书籍C	书籍D	书籍E	书籍F	书籍G
2021.1.1	17	6	8	24	13	13	18
2021.1.2	11	15	14	13	9	10	19
2021.1.3	10	8	12	13	8	3	7
2021.1.4	9	6	6	3	10	9	9
2021.1.5	4	10	13	8	12	10	17
2021.1.6	13	10	13	16	8	9	12
相关性系数：		-0.215108044	-0.272730395	0.821326501	0.044136741	0.323975798	0.140652396

图 3 相关性系数计算结果

用折线图来呈现，很直观看出来书籍 D 销售量总体上随着书籍 A 销售量的增大而增大，如图 4 所示。

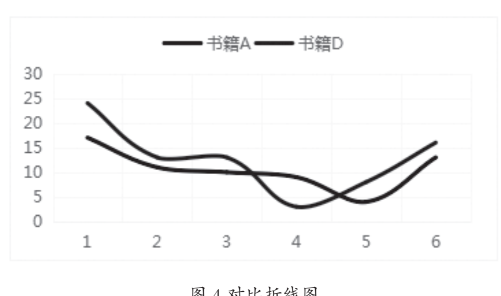


图 4 对比折线图

3.1.3 回归分析法

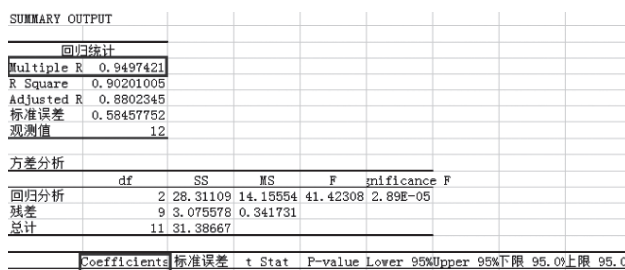
回归分析法用于确定因变量与自变量的关系，建立回归方程以表达相关性，可以预测因变量的未来变化。自变量与因变量可以不只一个，下面举例说明。

某运输公司为了制定优化的运输计划，为了确定承接的运货量希望能够预测每天司机的工作时间。经分析发现司机每天工作时间与运输距离、运输次数有关。特此采集了由 12 项运输活动组成的随机样本，并依据这些数据（如表 1 所示）构建二元线性回归方程。设时间为因变量，距离与次数为自变量，利用 Excel 回归分析工具对回归系数进行估算，得到的回归结果摘要如图 5 所示。复相关系数 R 为 0.9497，说明时间与运输距离及次数高度线性相关。依据计算结果，可以得到回归系数和回归方程：

$$y = -0.155 + 0.043x_1 + 0.544x_2$$

表 1 运输活动随机样本

时间	距离	次数	时间	距离	次数
9.3	160	4	6.2	128	2
4.8	80	3	7.4	120	3
8.9	161	4	6	105	4
6.5	101	2	7.6	145	3
4.2	80	2	6.1	145	1
5.2	96	3	9.0	150	5



SUMMARY OUTPUT

回归统计	
Multiple R	0.9497421
R Square	0.90201005
Adjusted R	0.8802345
标准误差	0.58457752
观测值	12

方差分析

	df	SS	MS	F	gnificance F
回归分析	2	28.31109	14.15554	41.42308	2.89E-05
残差	9	3.075578	0.341731		
总计	11	31.38667			

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	-0.1549379	0.780084	-0.19862	0.846977	-1.91961	1.609735	-1.91961	1.609735
X Variable	0.04314253	0.006295	6.853921	7.44E-05	0.028903	0.057382	0.028903	0.057382
X Variable	0.54434993	0.166343	3.272463	0.009645	0.168057	0.920643	0.168057	0.920643

图 5 回归结果摘要

回归系数 0.043 表示在固定次数的条件下，运输距离每增加 1 公里，行驶时间平均增加 0.043 小时；回归系数 0.573 表示在运输距离固定时，运输次数每增加 1 次，时间平均增加 0.544 小时。某司机某天运输 5 件货，最优路线总长为 150 公里，通过回归方程 y=–0.115+0.043 × 150+0.544 × 5 预测运输时间为 9.055 小时，与采样数据对比高度一致。

3.2 复杂模型

相对统计分析而言，较为复杂的经典算法模型有决策树、混沌理论、神经网络、蚁群算法及粒子群算法等，它们主要在人工智能、信息科学、控制论、机器学习等领域用于描述、分析、预测、优化、决策及控制等方面的应用。这些在前沿科学中多次被验证正确的复杂算法模型对于新媒体数据分析依然适用，如《深度神经网络视频新媒体短视频个性化推荐系统研究》[2] 等相关案例，下面以 BP 神经网络为例来描述算法模型的应用。

BP 神经网络全称为误差反向传播神经网络，在多层神经元网络中增加了用于连接权值的隐含层。BP 神经元的传递函数可采用线性函数或非线性函数，能逼近闭区间上的任何连续函数，致使某个神经元的输出可以有多值选择。BP 神经网络包括输入层、隐含层和输出层。输入输出向量间的非线性关系可由隐含层的非线性函数神经元加以描述。如图 6 所示的三层 BP 神经网络结构中，隐含层有 q 个神经元，输出层有 L 个神经元。输出层的误差反向传播可调整隐含层和输出层的权值。图 7 展现了运用 BP 神经网络对某服务系统网络流量负荷信息预报的情况。

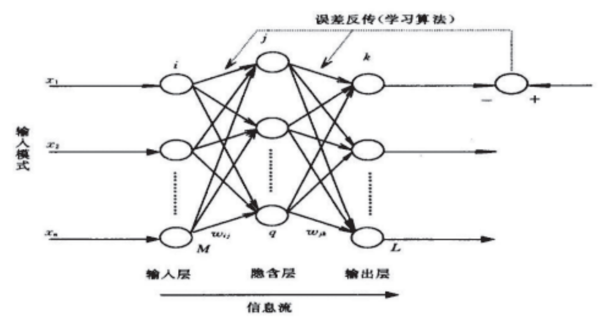


图 6 三层 BP 神经网络结构

样本日期	网络负荷	气象特征
2019-7-10	0.2452 0.1446 0.1314 0.2246 0.5532 0.6642 0.7015 0.6981 0.6821 0.6945 0.7549 0.8215 x1	
2019-7-11	0.2217 0.1581 0.1408 0.2304 0.5134 0.5312 0.6819 0.7125 0.7265 0.6847 0.7826 0.8325 y1 x2	0.2415 0.3027 0
2019-7-12	0.2525 0.1627 0.1507 0.2406 0.5502 0.5636 0.7501 0.7352 0.7459 0.7015 0.8064 0.8516 y2	0.2385 0.3125 0
2019-7-13	0.2016 0.1105 0.1243 0.1978 0.5021 0.5232 0.6819 0.6952 0.7015 0.6825 0.7825 0.7895 y3	0.2216 0.2701 1
2019-7-14	0.2115 0.1201 0.1312 0.2019 0.5532 0.5736 0.7029 0.7032 0.7189 0.7019 0.7965 0.8025 y4	0.2352 0.2506 0.5
2019-7-15	0.2335 0.1322 0.1534 0.2214 0.5623 0.5827 0.7198 0.7276 0.7359 0.7506 0.8092 0.8221 y5	Y 0.2542 0.3125 0
2019-7-16	0.2368 0.1432 0.1653 0.2205 0.5823 0.5971 0.7136 0.7129 0.7263 0.7513 0.8091 0.8217 y6	0.2601 0.3198 0
2019-7-17	0.2342 0.1368 0.1602 0.2131 0.5726 0.5822 0.7101 0.7098 0.7127 0.7121 0.7995 0.8216 y7	0.2579 0.3099 0
2019-7-18	0.2113 0.1212 0.1305 0.1819 0.4952 0.5312 0.6886 0.6898 0.6999 0.7323 0.7721 0.7956 y8 x9	0.2301 0.2867 0.5
2019-7-19	0.2005 0.1121 0.1207 0.1605 0.4556 0.5022 0.6553 0.6673 0.6798 0.7023 0.7521 0.7756 y9 x10	0.2234 0.2977 1
2019-7-20	0.2123 0.1257 0.1343 0.2079 0.5579 0.5716 0.7059 0.7145 0.7205 0.7401 0.8019 0.8136 K	0.2314 0.2977 0
2019-7-21	0.2119 0.1215 0.1621 0.2161 0.6171 0.6159 0.7155 0.7201 0.7243 0.7298 0.8179 0.8229 M	0.2317 0.2936 0

X

图 7 某服务系统网络流量负荷预报

负荷数据每天每隔 2 小时测得一次，还有最高气温、最低气温和天气特征（0 晴天、0.5 阴天、1 雨天），图 7 中数据已经归一化，共计 15 维输入向量（12 个网络负荷向量加 3 个气象特征向量），12 维输出向量（12 个网络负荷向量）。因此，网络输入层有 15 个神经元用 X 描述，输出层有 12 个神经元（激励函数为对数函数）用 Y 描述，隐藏层神经元个数设定为 31（激励函数为正切函数）。训练神经网络 Y=F（X，W），得到作为权值的 W 集合，再用 M=F（K，W）计算 M 得到预测结果。

3.3 大数据分析

大数据分析用来对具有海量规模的数据进行分析，其优势为快速流转数据、真实可靠。数字化新媒体下的数据分析技术能够成为人工智能的基础。[3] 与统计分析、算法模型不同的是大数据分析提供更加复杂稳定安全的数据分析系统体系，具备数据挖掘、预测性分析、数据仓库等功能。

在大数据应用系统中，Hadoop 作为处理大数据的分布式存储和计算框架，得到国内外大、中、小型企业的广泛应用。[4] 其核心包括可扩展、高容错、高性能的 HDFS 分布式文件系统和为海量数据的规则提取提供并行计算框架的 MapReduce。[5] 在 Hadoop 框架下实现 KNN 机器学习算法模型变得更加方便高效。

对一个样本，在特征空间中的 K 个最相邻的样本中的大多数属于某一个类别，则该样本也属于这个类别，并具有这个类别上样本的特性，这是 KNN 算法的核心思想。此方法在分类决策上只依据最邻近样本的类别来决定待分样本所属的类别，如图 8 展示了 KNN 分类原理。图中有两种类型的样本数据，一类是星形，另一类是椭圆形，圆形点是待分类数据。K=3 时，离圆形点最近的有 2 个椭圆形和 1 个星形，椭圆形占多数，那么待分类点属于椭圆形；K=6 时，离圆形点最近的有 2 个椭圆形和 4 个星形，星形占多数，那么待分类点属于星形。

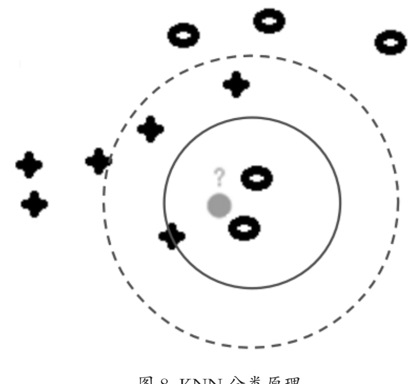


图 8 KNN 分类原理

当参与训练的数据海量时，由于单机内存和单机计算资源有限，导致传统 KNN 算法失效，又因每个训练样本不受其他训练样本影响，所以 KNN 能够被 MapReduce 实现。MapReduce 中最核心部件是 map 与 reduce 函数，map 将大任务划分为若干小任务，小任务能够同时运行，再通过 reduce 将多个小任务的结果汇总起来。对于 KNN，训练数据量大可将训练数据分布式存储读入 map 中，在 map 中每输入一个训练样本就计算它和所有测试数据的距离并传到 reduce 中，然后 reduce 将同一个测试数据的距离合并然后排序计数得到测试样本的类标识并输出。由于 Hadoop 的 MapReduce 计算框架遵循 key–value（键值对）原则，map 与 reduce 函数设计如表 2 所示。

表 2　map 与 reduce 函数设计

函数名称	任务	输入数据形式	输出数据形式
map	读取训练集，计算测试样本与每个训练样本之间的相似度	<key，value> 键值对 key：训练样本行号 value：训练样本（假设每行对应一个训练样本）	<key，value> 键值对 key：测试样本 ID value：< 训练样本的类标识，相似度值 >
reduce	找出 K 个近邻，计算多数类的类别，并将其赋予测试样本	<key，value> 键值对 key：测试样本 ID value：iterator< 训练样本的类标识，相似度值 >	<key，value> 键值对 key：测试样本 ID value：测试样本的类标识

以上是利用 Hadoop 实现 KNN 算法的应用设计，达到简洁高效、并行处理完成分类任务的目的，此种方法在新媒体中适合于文本分类、舆情分析、舆情预测等关键问题的解决。

结语

综上所述，文中提及的新媒体数据分析方法均有各自特征，统计分析法对历史数据的完整性和准确性要求较高，分析步骤简单、容易掌控；复杂模型如 BP 神经网络具有超强的学习能力与容错能力，能够处理复杂的非线性关系；大数据分析提出了针对海量数据的、具备架构体系规模的解决方案，兼顾规范性的同时不失处理数据的灵活性，由专业人士完成数据分析过程。为此，需要依据现实问题的实际情况与新媒体运营企业所拥有的资源，合理发挥各自长处，确保新媒体数据分析结果的有效性。媒

参考文献

[1] 段峰峰 . 新媒体数据分析与应用 [M]. 北京：人民邮电出版社，2020：100.

[2] 高晨峰 . 深度神经网络视频新媒体短视频个性化推荐系统研究 [J]. 卫星电视，2019（5）：16-20.

[3] 段汝林 . 数字化新媒体下数据分析技术的应用研究 [J]. 教育现代化，2016（36）：215-216.

[4] 余明辉，张良均 .Hadoop 大数据开发基础 [M]. 北京：人民邮电出版社，2018：1.

[5] 刘春阳，张学龙，刘丽军等 .Hadoop 大数据开发 [M]. 北京：中国水利水电出版社，2018：12.

作者简介：王楷鑫（2001-），女，吉林，研究方向：网络与新媒体；王双立（1972-），男，吉林省吉林市，副教授，研究方向：数据分析与应用、智能科学与技术。

（责任编辑：张晓婧）